Children, Humanoid Robots and Caregivers

Artur Arsenio
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
arsenio@csail.mit.edu

This paper presents developmental learning on a humanoid robot from human-robot interactions. We consider in particular teaching humanoids as children during the child's Separation and Individuation developmental phase (Mahler, 1979). Cognitive development during this phase is characterized both by the child's dependence on her mother for learning while becoming awareness of her own individuality, and by self-exploration of her physical surroundings. We propose a learning framework for a humanoid robot inspired on such cognitive development.

1. Introduction

This paper describes work which takes inspiration on Mahler's child development theory (Mahler, 1979). Special emphasis is put on the child's *Separation and Individuation* developmental phase (Mahler, 1979) – during which the child eventually separates from his mother bound and embraces the external world. Mahler's theory has influences by movements such as the Ego's Developmental Psychology and Experimental Psychology, from Freud, Piaget and others. According to her theory, the normal development of a child during the Separation and Individuation phase is divided into four sub-phases, following the *Epigenetic Principle* – as each stage progresses, it sets the foundation for the next stages:

Differentiation (5-9 months) The first sub-phase, marked by a decrease of the infant's total dependency on his mother as the former crawls further away. The infant starts to realize his own individuality and separateness due to the development of the entire sensory apparatus and therefore a growing awareness

Practicing (9,10-18 months) Sub-phase characterized by the child's active locomotion and exploration of his surroundings, together with the narcissist exploration of his own functions and body

Re-approximation (15-24 months) Child has an egocentric view of the world during this phase, in which he also approximates again to his mother. World expands as new viewing angles are available from the child's erect walking

Individuality and Object Constancy (24-36 months) Defined by the consolidation of individuality, and a clear separation between objects and himself. Towards the end, the child becomes aware of object constancy.

Therefore, during the Separation and Individuation phase, the child learns to recognize himself as an individual, and his mirror image as belonging to himself. He learns also about its surrounding world structure - about probable places to find familiar objects (such as toys) or furniture items. In addition, he starts to identify scenes - such as his own bedroom and living-room. And children become increasingly aware (and curious) of the outside world (Lacerda et al., 2000). This paper describes the implementation of these cognitive milestones on the humanoid robot Cog, placing special emphasis on developmental object perception (Johnson, 2002) during the Separation and Individuation stage.

The child's mother plays an essential role (Gonzalez-mena and Widmeyer, 1997) in guiding the child through this learning process. Aiming at teaching humanoid robots as children during this stage, the child's mother role will be attributed to a human tutor/caregiver. Therefore, a human-centered approach is presented to facilitate the robot's perception and learning, while showing the benefits that result from introducing humans in the robot's learning loop.

2. Learning on the Autistic and Symbiotic Phases

This section reviews shortly the methodology we developed for robot interactions motivated by infant's simple learning mechanisms in Mahler's autistic and Symbiotic developmental phases, which antecede the Separation and Individuation phase. In the autistic phase (from birth to 4 weeks old), the newborn is most of the time in a sleeping state, awakening to eat or satisfy other necessities (Mahler, 1979, Muir and Slater, 2000). His motor skills consist mainly of primitive reflexes until the end of this phase (Michel and Moore, 1995). Towards the Symbiotic phase (until 4-5 months), the

infant's attention is often dropped to objects under oscillatory motions, or to abrupt changes of motion, such as throwing an object. Baby toys are often used in a repetitive manner – consider rattles, car/hammer toys, etc. This repetition can potentially aid a robot to perceive these objects robustly. Playing with toys might also involve discontinuous motions (for instance, grabbing a rattle results in a sharp velocity discontinuity upon contact).

This motivated the design of algorithms which implement the detection of events with such characteristics. Moving image regions that change velocity either periodically, or abruptly under contact produce visual event candidates. These algorithms, which are presented in detail by (Arsenio, 2003), identify such events at multiple spatial/frequency resolutions. Object Segmentation, a fundamental problem in computer vision, is then dealt with by detecting and interpreting natural human/robot task behavior from discontinuous events – such as tapping, waving, shaking, poking, grabbing/dropping or throwing objects – or from periodic events – such as waving or shaking an object (Arsenio, 2003, Arsenio et al., 2003).

An active segmentation technique developed recently (Fitzpatrick, 2003) relies on poking objects with a robot actuator. This strategy operates on first-person perspectives of the world: the robot watching its own motion. However, it is not suitable for segmenting objects based on external cues. We would like therefore to transfer skills from humans to the robot, so that external information can be incorporated to enable autonomous acquisition of knowledge by the robot, by exploiting shared world perspectives between a cooperative human and the robot. Such developmental approach for skill transfer is presented by (Arsenio, 2004a). By observing a human interacting with objects (for instance, waving or poking them), the robot builds Hybrid Markov Models that model the task, and is then able to act by itself on (un)known objects to segregate them from the background, as shown in Figure 1.

The child's *Separation and Individuation* phase (Mahler, 1979) is marked by the separation of the child from his mother as a different individual. However, the child still relies heavily on help provided by his mother to understand the world and even himself through this developmental phase (Gonzalez-mena and Widmeyer, 1997). Indeed, the child is part of a structured world that includes the immediate emotional (for robot emotions, not covered by this paper, see (Breazeal, 2000)), social and physical surroundings (Michel and Moore, 1995). In the following sections, social help from a human tutor will be used to guide the robot through learning about its physical surroundings. In particular, this helping hand will assist the robot to correlate data

Figure 1: a) Object segmentations extracted from human created events b) Object segmentations extracted from robot created events c) robot executes a simple learned task (waving), and associates the sound to the movement of its own body d) top: sequence of images extracted from a poking event; bottom: object and actuator segmentation from a poking event created by the robot.

among its own senses (Section 3.); to control and integrate situational cues from its surrounding world (Section 4.); and to learn about out-of-reach objects and the different representations in which they might appear (Section 5.). Special emphasis will therefore be placed on social learning along a child's physical topological spaces, as shown in Figure 2.

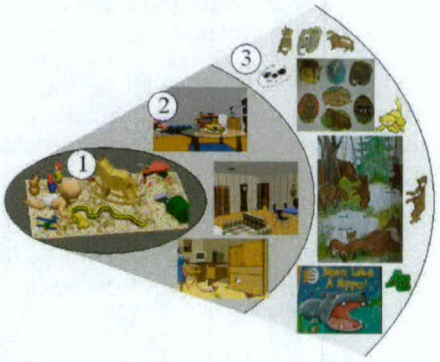

Figure 2: Developmental learning during the child's *Separation and Individuation* phase will be described along three different topological spaces: 1) the robot's personal space, consisting of itself and familiar, manipulable objects (Section 3.); 2) its living space, such as a bedroom or living room (Section 4.); and 3) its outside, unreachable world, such as the image of a bear on a forest (Section 5.).

3. Learning about Objects and Itself

This section describes a strategy for a robot to associate data from several sources: from its own senses, to better perceive both itself and objects with which it interacts, and from its senses and information stored on the world or on the robot's memory.

Information is gathered from a human tutor creating rhythmic actions, facilitating this way the robot's perception. This is motivated from a child's mother role in helping the child to learn about objects and its own body – by tapping or waving a toy or a child's body part (such as a hand) while annunciating the name associated to it, or by performing educational activities, such as drawing or painting.

3.1 Cross-Modal Data Association

Cross-modal data association from the robot's own senses is briefly described hereafter (for details see (Fitzpatrick and Arsenio, 2004)). This is an important capability for a humanoid robot, so that it can better perceive itself and objects with which it interacts. In children, such capability appears with the development of the sensory apparatus on the first differentiation sub-phase.

Due to physical constraints, the set of sounds that can be generated by manipulating an object is often quite small. For toys which are suited to one specific kind of manipulation – as rattles encourage shaking – there is even more structure to the sound they generate (Fitzpatrick and Arsenio, 2004). When sound is produced through motion for such objects the audio signal is highly correlated both with the motion of the object and the tools' identity. Therefore, the spatial trajectory can be applied to extract visual and audio features – patches of pixels, and sound frequency bands – that are associated with the object (see Figure 3), which enables the robot to map the visual appearance of objects manipulated by humans or itself to the sound they produce.

Proprioceptive data is a sensorial modality very important to control the mechanical device, as well as to provide workspace information (such as the robot's gaze direction). But it is also very useful to infer identity about the robotic *self* (Fitzpatrick and Arsenio, 2004) (for instance, by having the robot recognize itself on a mirror). Children become able to self-recognize their image on a mirror during the practicing sub-phase, which marks an important developmental step towards the child individuality. On a humanoid robot, large correlations of a particular robot's limb with data from other sensorial inputs indicates a link between such sensing modality to that moving body part (which generated a sound, or which corresponds to a given visual template, as shown in Figure 4). Therefore, the binding algorithm was extended to account for proprioceptive data, which is matched to both visual and audio signals. Such an approach enables not only the identification of the robot's own acoustic rhythms, but also the visual recognition of the robot's mirror image, as shown in Figure 4 (this is an important milestone on the development of a child's theory of mind (Baron-Cohen, 1995)).

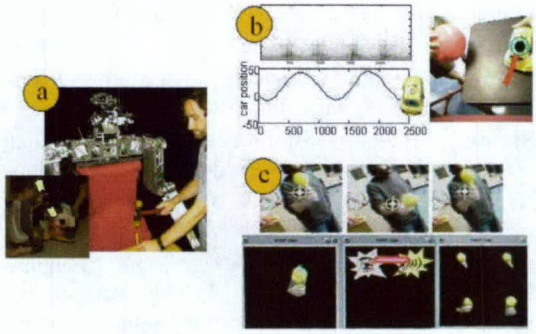

Figure 3: a) A child and a human playing with a hammer, which bangs in a table, producing a distinctive audio signal. b) A human moves a car repetitively forward/backward producing sound in each direction, which is matched to the visual trajectory. The sound energy has two peaks per visual period, since the sound of rolling is loudest during the two moments of high velocity motion between turning points in the car's trajectory (because of mechanical rubbing). c) top: tracking an oscillatory instrument; down: image of object segmentation and display of a detected visual/sound matching.

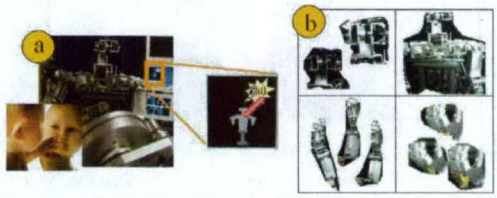

Figure 4: a) Child and robot looking at a mirror, associating their image to their body (image of robot/sound association shown amplified for the robot) b) Visual segmentations organized according to the robot's body parts for which they were matched.

3.2 Object Recognition

Sensorial data strongly correlated to proprioceptive data is therefore labelled with the correspondent robot's body part. However, it is necessary to develop a recognition scheme for objects other than robot's body parts, which enables object recognition under different contexts.

The object recognition algorithm consists of three independent algorithms. Each recognizer operates along orthogonal directions to the others over the input space (Arsenio, 2004b). This approach offers the possibility of priming specific information – which will be shown a property of paramount importance – such as searching for a specific object feature (color, shape or luminance) independently of the others. For instance, the recognizer may focus the search on a specific color or sets of colors, or look into both desired shapes and luminance (Arsenio, 2004b):

Color. Input features consist of groups of connected regions with similar color

Luminance. Input space consists of groups of connected regions with similar luminance

Shape. A Hough transform is applied to a contour image (from a Canny edge detector). Line orientation is determined using Sobel masks. Pairs of oriented lines are then used as input features

Geometric hashing is a rather useful technique for high-speed performance. In this method, invariants (or quasi-invariants) are computed from training data in model images, and then stored in hash tables. Recognition consists of accessing and counting the contents of hash buckets. An Adaptive Hash table (Arsenio, 2004b) (a hash table with variable-size buckets) was implemented to store affine color, luminance and shape invariants (which are view-independent for small perspective deformations). Figure 5 shows results for each of these input spaces, while experimental results for real objects will be shown in the next sections.

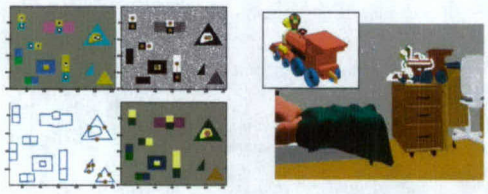

Figure 5: (left) Conjunction searches; Top row, from left to right: normalized color buckets for the original image, with results for a yellow-green query superimposed; and Luminance buckets of the original image, together with query results for a dark-light object; Bottom row: Search for triangles (conjunction of three oriented lines); and the target identification (a conjunction of features) among distracters. (right) Object recognition and location. The train appears under a perspective transformation in a computer generated bedroom. Scene lines matched to the object are outlined.

3.3 Learning from Educational Activities

A common pattern of early human-child interactive communication is through activities that stimulate the child's brain, such as drawing or painting. Children on the practicing sub-phase of development, and older, are able to extract information from such activities while they are being performed on-line. This capability motivated the implementation of three parallel processes which receive input data from three different sources: from an attentional tracker (Fitzpatrick, 2003), which tracks the attentional focus and is attracted to a new salient stimulus; from a multi-target tracking algorithm, implemented to track simultaneously multiple targets; and from an algorithm that selectively attends to the human actuator for the extraction of periodic signals from the trajectory of oscillating skin blobs.

Whenever a repetitive trajectory is detected from any of these parallel processes, it is partitioned into a collection of trajectories, being each element of such collection described by the trajectory points between two zero velocity points with equal sign on a neighborhood (similarly to the partitioning process described in (Fitzpatrick and Arsenio, 2004)). As shown in Figure 6, the object recognition algorithm is then applied to extract correlations between these sensorial signals perceived from the world and geometric shapes present in such world, or on the robot object database, as follows:

1. Each partition of the repetitive trajectory is mapped into a set of oriented lines by application of the Hough transform.

2. By applying the recognition scheme previously described, trajectory lines are matched to oriented edge lines (from a Canny detector) on

   (a) a stationary background,

   (b) objects stored in the robot's object recognition database.

This way, the robot learns object properties not only through cross-modal data correlations, but also by correlating human gestures and information stored in the world structure (such as objects with a geometric shape) or on its own database.

On children, such capabilities evolve according to the epigenetic principle as they start to move around on their physical surroundings, learning about its structure. This occurs mainly during the practicing developmental sub-phase, and towards the reapproximation phase the child gets a completely new view of the world from erect walking.

4. Learning the World Structure of the Robot's Physical Surroundings

Autonomous agents, such as robots and humans, are situated in a dynamic world, full of information stored on its own structure. For instance, the probability of a chair being located in front of a table is much bigger than that of being located on the ceiling. A robot should place an object where it can easily find it - if one places a book on the fridge, he will hardly find it later!

This dual perspective on object recognition is an important milestone for children - not only to be able to infer the presence of objects based on the scene context, but also to be able to determine where objects should be stored based on the probability of finding them on that place later on.

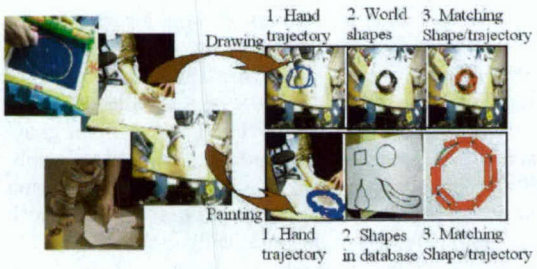

Figure 6: Learning activities, such as drawing on paper or boards. Shows a human painting a black circle on a sheet of paper with a ink can. The circle is painted multiple times. The hand trajectory is shown, together with edge lines on the background image matched to such trajectory. It also shows a human drawing a circle on a sheet of paper with a pen, which is matched into a circle drawn previously and stored in the robot's database.

Therefore, a statistical framework was developed to capture knowledge stored in the robot's surrounding world. This framework consists of: 1) learning 3D scenes from cues provided by a human actor; and 2) learning the spatial configuration of objects within a scene.

4.1 Learning about Scenes

The world structural information should be exploited in an active manner. A significant amount of contextual information may be extracted from a periodically moving actuator – most often such motions are from interactions with objects of interest – which can be framed as the problem of estimating $p(o_n|v_{B_{\vec{p},\epsilon}}, act^{per}_{\vec{p},S})$, the probability of finding object n given a set of local, stationary features v on a neighborhood ball B of radius ϵ centered on location p, and a periodic actuator on such neighborhood with trajectory points in the set $S \subseteq B$. The *Segmentation from Demonstration* method which will be described in Section 5. solves such problem.

The environment surrounding the robot also provides additional structure that can be learned through supervised learning techniques. Hence, scenes will be defined as a collection of objects with an uncertain geometric configuration, each object being within a minimum distance from at least another object in the scene. Figure 7 presents statistical results for segmentations of several furniture items on a scene. Scene descriptions are built by mapping all information about objects (mainly furniture) into egocentric coordinates. Figure 7 also shows both the reconstruction of the visual appearance of a scene in the robot's lab and a coarse depth image for such scene.

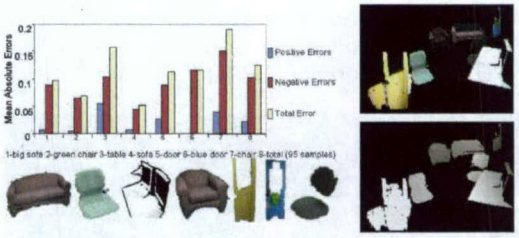

Figure 7: (left) Segmentation error analysis for furniture items on a scene – segmentation samples are also shown (right) Furniture image segmentations– on top – and depth map – bottom – for a scene in Cog's room. Depth maps are extracted by an active, embodied approach that relies on a human to actively change the context of a scene, so that the human arm diameter is used as a reference for extracting relative monocular depth.

4.2 Learning about Objects in Scenes

Children need to be able not only to built environment descriptions for safety locomotion, but also to learn the relative probability distribution of objects in a scene – for instance, books are often found on top of shelves. Therefore, the scene context puts a very important constraint on the type of places in which a certain object might be found. From a humanoid point of view, contextual selection of the attentional focus is very important both to constrain the search space for locating objects (optimizes computational resources) and also to determine common places on a scene to drop or store objects such as tools or toys.

Given the image of an object, its meaning is often a function of the surrounding context. Context cues are useful to remove such ambiguity. Ideally, contextual features should incorporate the functional constraints faced by people, objects or even scenes (eg. people cannot fly and offices have doors). Therefore, functionality plays a more important role than more ambiguous and variable features (such as color, which selection might depend on human taste). Functionality constraints have been previously exploited for multi-modal association (Fitzpatrick and Arsenio, 2004) and for determining function from motion (Duric et al., 1995), just to name a few applications.

As such, texture properties seem appropriate, which led to the selection of Wavelets (Strang and Nguyen, 1996) as contextual features, since they are much faster to compute than Gabor filters and provide a more compact representation. Input monochrome images are transformed using a Daubechies-4 wavelet tree, along 5 depth scales. The input is represented by $v(\vec{p}) = \{v_k(x,y), k = 1,\ldots,N\}$, with N=15. Each wavelet component is down-sampled to a 8×8 image, so that $\bar{v}(x,y)$ has dimension 960.

Figure 8 shows image reconstruction from the sets of features $v(\vec{p})$ (also denoted image sketch or holistic representation (Torralba, 2003)).

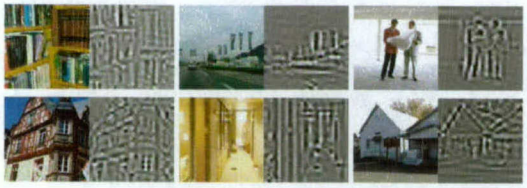

Figure 8: Reconstruction of the original image (by the inverse Wavelet transform). As suggested by (Torralba, 2003), this corresponds to an holistic representation of the scene.

The dimensionality problem is reduced to become tractable by applying Principal Component Analysis (PCA). The image features from the wavelet transformation $\bar{v}(\vec{p})$ are decomposed into the basis functions provided by the PCA, encoding the main spectral characteristics of a scene with a coarse description of its spatial arrangement. The decomposition coefficients are obtained by projecting the image features $v_k(\vec{p})$ into the principal components $\vec{c} = \{c_i, i = 1, \ldots, D\}$ ($\vec{c}$ denotes the resulting D-dimensional input vector, used thereafter as input context features). These coefficients can be viewed as a scene's holistic representation since all the regions of the image contribute to all the coefficients, as objects are not encoded individually. The effect of neglecting local features is reduced by mapping the foveal camera (which grabs data for the object recognition scheme based on local features described in Section 3.2) into the image from the wide field of view camera, so that the weight of the local features is strongly attenuated. The position vector $\vec{p}$ is thus given in wide field of view retinal coordinates.

The output space is defined by the 6-dimensional vector $\vec{x} = (\vec{q}, d, \vec{s}, \phi)$, where $\vec{q}$ is the object's centroid – a 2-dimensional position vector in wide-field of view retinal coordinates, d is the object's depth, $\vec{s} = (w, h)$ is a vector containing the principal components of the ellipse that models the 2D retinal size of the object, and ϕ is the orientation of such ellipse.

A method based on a weighted mixture of gaussians was applied to find interesting places where to put a bounded number of local kernels that can model large neighborhoods. Therefore, given the context $\vec{c}$, one needs to evaluate the PDF $p(\vec{x}|o_n, \vec{c})$ from a mixture of m (spherical) Gaussians (Gershenfeld, 1999):

$$p(\vec{x}, \vec{c}|o_n) = \sum_{m=1}^{M} b_{mn} G_x(\vec{x}, \vec{\eta}_{mn}, X_{mn}) G_c(\vec{c}, \vec{\mu}_{mn}, C_{mn})$$

The mean $\vec{\eta}_{mn}$ of the Gaussian G_x is a function that depends on $\vec{c}$ and on a set of parameters β_{mn}. A locally affine model was chosen for the mean: $\beta_{m,n} = (\vec{a}_{m,n}, A_{i,n})$: $\vec{\eta}_{m,n} = \vec{a}_{m,n} + A^T \vec{c}$. The EM algorithm is then used to learn the cluster parameters (see (Gershenfeld, 1999) for a detailed description of the EM algorithm). The number M of gaussian clusters is selected in order to maximize the join likelihood of the data. An agglomerative clustering approach based on the minimum description length was implemented to automatically estimate M.

Figure 9 presents results for selection of the attentional focus for several furniture objects. However, there is a lot of information that cannot be extracted from scenes familiar to a robot (real whales are not common in humanoid research labs). But such information from the robot's outside world can be transmitted to the robot by a human tutor using books.

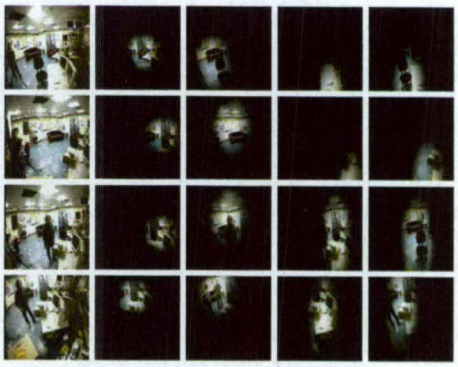

Figure 9: Samples of scene images are shown on the first column. The next four columns show probable locations based on context for the smaller sofa, the bigger sofa, the table and the chair, respectively. Notice that, even if the object is not visible or present, the system estimates the places at which there is a high probability of finding such object. Two such examples are shown for the chair – no matter the viewing angle, chairs are predicted to appear in front of the table. It is also shown that occlusion by humans do not change significantly the global context.

5. Learning about the Outside World through Books

Children's learning is often aided by the use of audio-visuals, and especially books, from social interactions with their mother or caregiver during the developmental sub-phases of re-approximation and individual consolidation, and afterwards. Indeed, humans often paint, draw or just read books to children during their childhood. Books are indeed a useful tool to teach robots different object representations and to communicate properties of unknown objects to them.

Learning aids are also often used by human caregivers to introduce the child to a diverse set of (in)animate objects, exposing the latter to an out-

side world of colors, forms, shapes and contrasts, that otherwise might not be available to a child (such as images of whales or, as shown by Figure 10, images of cows). Since these learning aids help to expand the child's knowledge of the world, they are a potentially useful tool for introducing new informative percepts to a robot.

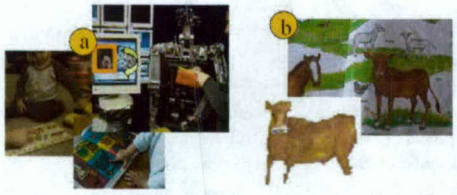

Figure 10: a) Child and human learning from a book. A car template extracted by the robot during an experiment is shown outlined by a square. b) Extraction of a cow's template image from a book.

The strategies which enable the robot to learn from books rely heavily in human-robot interactions. It is essential to have a human in the loop to introduce objects from a book to the robot (as a human mother/caregiver does to a child), by tapping on their book's representations. Segmentation by demonstration – a human aided object segmentation algorithm – segments an object's image from book pages (or a furniture item from the background), as follows (Arsenio, 2004b):

1. A color segmentation algorithm is applied to a stationary image
2. A human actor waves the arm/hand/finger on top of the object to be segmented
3. The motion of skin-tone pixels is tracked over a time interval (using the Lucas-Kanade algorithm). The energy per frequency content – using Short-Time Fourier Transform (STFT) – is determined for each point's trajectory
4. Periodic, skin-tone points are grouped together into the arm mask (Arsenio, 2003).
5. The trajectory of the arm's endpoint describes an algebraic variety over N^2 (N represents the set of natural numbers). The target object's template is given by the union of all bounded subsets (the color regions of the stationary image) which intersect this variety

This grouping works by having trajectory points being used as seed pixels. The algorithm (see Figure 11) fills the regions of the color segmented image whose pixel values are closer to the seed pixel values, using a 8-connectivity strategy.

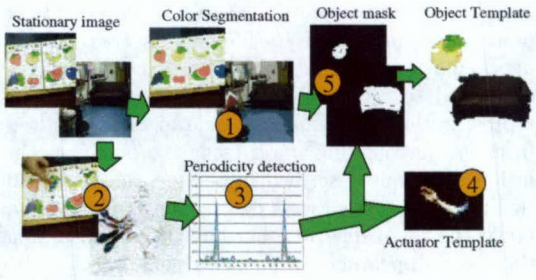

Figure 11: The actuator's trajectory is used to extract the object's color clusters.

Therefore, points taken from waving are used to both select and group a set of segmented regions into the full object. This strategy segments objects that cannot be moved independently, such as objects printed in a book, or heavy, stationary objects such as a table or a sofa. This scheme was successfully applied to extract templates for animals (many of which might not be visually accessible to the child from sources other than learning aids), furniture items, musical instruments, fruits, clothes, geometric shapes and other elements from books, under varying light conditions (as shown in Figure 12).

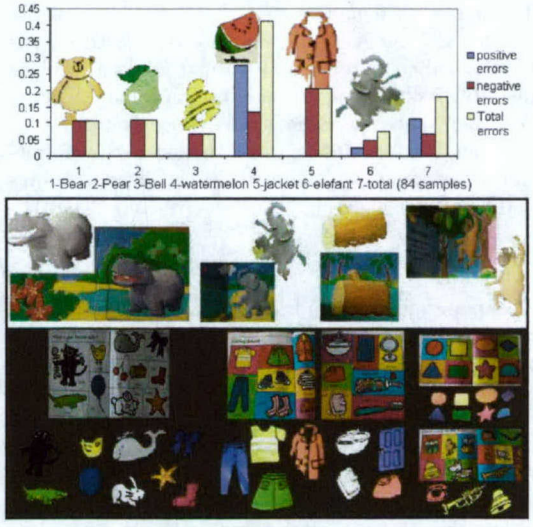

Figure 12: Statistical analysis for object segmentation from books. Templates for several categories of objects (for which a representative sample is shown), were extracted from a collection of children books.

5.1 *Matching Multiple Representations*

Object representations acquired from a book are inserted into a database, so that they become avail-

able for future recognition tasks. However, object descriptions may came in different formats - drawings, paintings, photos, etc. Hence, methods were developed to establish the link between an object representation in a book and *real* objects recognized from the surrounding world using the object recognition technique described in Section 3.2, as shown by Figure 13. Except for a description contained in a book, the robot had no other knowledge concerning the visual appearance or shape of such object.

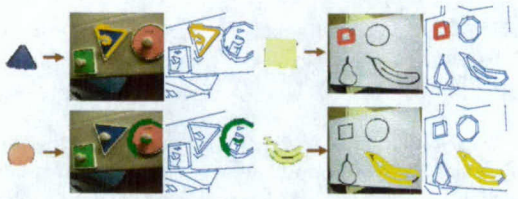

Figure 13: (left) Geometric shapes recognized using the descriptions from a book triangle–top– and circle–bottom. The recognition from chrominance features is trivial – objects have a single, identical color (right) Recognition of geometric, manual drawings from the description of objects learned using books.

Additional possibilities include linking different object descriptions in a book, such as a drawing, as demonstrated by the results presented in Figure 13. These results demonstrated the advantages of object recognition over independent input features: the topological color regions of a square drawn in black ink are easily distinguished from a yellow square. But they share the same geometric contours.

Other feasible descriptions to which this framework was applied include paintings, prints, photos and computer generated objects (Arsenio, 2004b).

6. Conclusions

This paper described a framework for developmental object perception and learning on a humanoid robot, which aims at teaching humanoids as children. We described algorithms to learn about the robot's self appearance and its surrounding world. The epigenetic principle established the foundations for these algorithms. Learning about new object representations was possible after some knowledge about the object (for example, from a book) was actively introduced by a human actor. The robot learned about its surrounding world by first building scene descriptions of world structures. Such descriptions then generated training data which enabled contextual selection of the attentional focus – to find regions on the visual field where there is a high probability of finding an object. We also shown learning from educational activities, such as drawing, enabled by previous storage of information concerning object shapes.

Children development is indeed a rich source of inspiration towards cognitive development on humanoid robots. But to achieve such an endeavor the mother's (and the child's caregiver) role should not be neglected – robots will also benefit from having a human helping hand that guides them through the learning process.

References

Arsenio, A. (2003). Embodied vision - perceiving objects from actions. *IEEE International Workshop on Human-Robot Interactive Communication.*

Arsenio, A. (2004a). *Developmental Learning on a Humanoid Robot.* accepted for the International Joint Conference on Neural Networks.

Arsenio, A. (2004b). Teaching a humanoid robot from books. In *International Symposium on Robotics.*

Arsenio, A., Fitzpatrick, P., Kemp, C. C., and Metta, G. (2003). The whole world in your hand: Active and interactive segmentation. Proceedings of the Third International Workshop on Epigenetic Robotics.

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press.

Breazeal, C. (2000). *Sociable Machines: Expressive Social Exchange Between Humans and Robots.* PhD thesis, MIT, Cambridge, MA.

Duric, Z., Fayman, J., and Rivlin, E. (1995). Recognizing functionality. In *Proc. International Symposium on Computer Vision.*

Fitzpatrick, P. (2003). *From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot.* PhD thesis, MIT, Cambridge, MA.

Fitzpatrick, P. and Arsenio, A. (2004). *Feel the beat: using cross-modal rhythm to integrate robot perception.* Submitted to fourth International Workshop on Epigenetic Robotics.

Gershenfeld, N. (1999). *The nature of mathematical modeling.* Cambridge university press.

Gonzalez-mena, J. and Widmeyer, D. (1997). *Infants, Toddlers, and Caregivers.* Mountain View.

Johnson, S. (2002). *Development of object perception,* pages 392–399. Nadel, L and Goldstone, R., (Eds.) Encyc. Cognitive Science. Macmillan, London.

Lacerda, F., Hofsten, C., and Heimann, M. (2000). *Emerging Cognitive Abilities in Early Infancy.* Erlbaum.

Mahler, M. (1979). *The Selected Papers of Margaret S. Mahler : Separation-Individuation,* volume II. NY.

Michel, G. and Moore, C. (1995). *Developmental Psychobiology: An Interdisciplinary Science.* MITPress.

Muir, D. and Slater, A. (2000). *Infant Development: The essential readings.* Essential readings in Developmental Psychology.

Strang, G. and Nguyen, T. (1996). *Wavelets and Filter Banks.* Wellesley-Cambridge Press.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision,* pages 153–167.